

DOCUMENT RESUME

ED 282 398

FL 016 678

AUTHOR Mason, Victor W.
TITLE Preparing Tests of Reading Comprehension.
PUB DATE [86]
NOTE 23p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communicative Competence (Languages); *English
(Second Language); *English for Academic Purposes;
Higher Education; Item Sampling; Language
Proficiency; *Language Tests; *Reading Skills;
*Reading Tests; *Test Construction; Test Items; Test
Use; Test Validity; Time Factors (Learning)

ABSTRACT

Reading skills are crucial to students learning and using English as a second language for academic purposes. Teachers can construct valid reading tests if they approach the task with care and focus on the test's ability to measure construct rather than face validity. In reading tests, the crucial elements of test design affecting validity are (1) the content of the reading tasks, (2) the length of the stimulus material, (3) the question format, (4) the time allotted for completion of the task, and (5) the amount of information obtained about the students' reading ability. Good reading tests should also show item sampling reliability, contain items whose formats require productive language, and promote scoring reliability. Good tests assess student proficiency and progress accurately and also give students confidence in their effectiveness, comprehensiveness, stringency, and accuracy. (MSE)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Preparing Tests of Reading Comprehension

ABSTRACT

The paper emphasizes the exceptional importance to students of reading ability in the typical EFL or ESL program at an academic institution. Concomitant importance thus attaches to the ability 1) of instructors both to teach reading effectively and to design and write good reading tests; and 2) of teacher-preparation programs to augment the test-development skills and knowledge of degree candidates.

Discussion focuses on those design features of reading tests promoting or inhibiting their validity. It is stressed that score validity, not so-called face validity, needs to be the test writer's and test administrator's main concern. The paper concludes that the test format best promoting the construct and criterion-related validity of well-designed and -written reading measures is the 4-option multiple-choice format.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

V. Mason

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Preparing Tests of Reading Comprehension

I. General Introductory Considerations

Needs analyses, whether formal or informal, consistently point out that reading ability is one of the most important skills for university-level EFL students to develop. Whether enrolled in EAP or ESP courses, students depend fundamentally on their ability to comprehend progressively more complex, sophisticated and lengthy text(s) in order to become linguistically more proficient---and to pass their courses with satisfactory marks.

Course statements of linguistic and behavioral objectives thus typically place great importance on developing EFL students into ever more proficient and fluent readers. And teachers, concerned that they validly test what they teach, devote no little time during a course to the preparation of reading tasks, to assess the extent of their students' progress in this vital macro-skill area.

However, teachers are often uncertain about the "best" way to test reading comprehension of EFL students in an academic program. They may disagree on the nature of the stimulus material---i.e., the reading passage(s) presented---or on the kinds of questions to set; or both. They may be unsure, for several obvious reasons. First, as Woodford notes, many teacher-preparation

programs traditionally have not emphasized sufficiently the development of measurement and evaluation ability of their degree candidates (Woodford 99). Moreover, when such programs do not provide adequate practical application of instructional theory in classrooms of real EFL or ESL students (Orem 2), attention to testing in a course may be as removed from reality as is discussion of teaching theory and methodology. For it is as vital to a novice instructor to have extended, supervised "hands on" experience teaching and testing students as it is for a student doctor, dentist or veterinarian to work on a continuing basis with real patients.

As a consequence, a great many English language teaching (ELT) practitioners are forced to learn much, if not most, of what they know as veteran teachers and testers from practical, on-the-job experience. If that experience---no matter how long---has been insufficient, the complexities of measurement and evaluation may produce feelings of insecurity in ELT practitioners, ranging from guilt (Alderson and Hughes 1981), to defensiveness (Woodford 99), to fear (Cohen 1980). Palmer (1983) has pointed out that the extent of professional preparation and training in teaching that EFL teachers receive often "seems not to correlate with their professional success." The same may often be said of their preparation, and subsequent experience, in testing work. Thus, the applicability of Henning's observations and injunctions, about common ELT

testing errors to avoid, remains extremely widespread (Henning 1982).

One of the main purposes of this article is to point out that, while testing is not an uncomplicated subject, any individual of average ability can become reasonably proficient in his or her testing skills. The requisite qualities are: 1) common sense in trying to solve problems; 2) the readiness to work hard---by spending enough time---in dealing with test-related tasks; and 3) the willingness regularly to consult references specifically written for teachers on practical testing issues.

Finally, and not least importantly, a major cause of difficulties in measurement-related matters is that this subject tends to be one of the most emotional of all ELT topics. At the level of theory, the literature repeatedly demonstrates the amount of heat that disagreements over this subject generate among the testing authorities. Similarly, discussion at the level of the teaching program often becomes just as charged. This paper hopes to help reconcile differences in testing views (Nakhoul 1985) and to clarify issues related to one of the most difficult types of ELT testing tasks required of practitioners, the devising of good measures of reading comprehension. It does so in the belief that the process of that test-development can be carried out in programs anywhere in a manner which is both efficient and harmonious. It

does so also on the assumption that as ELT staffers increase their understanding of basic testing principles, they become more confident in their testing abilities and find that the topic is no longer an intimidating one but rather one of the most interesting and challenging aspects of their professional careers.

II. The Special Relevance of Reading Tests in ELT Programs

A. The Importance of Reading in the Curriculum

In the typical ELT program, no skill is more important to develop in students than reading ability, as needs analyses repeatedly emphasize. In most EFL situations, reading is probably the skill upon which students most depend to make meaningful progress in English and in their FL-related academic studies. Even in ESL settings, with a greater balance usually required in proficiency among the four macro-skills, again it is plausible to surmise that reading, if not pre-eminent in the overall curriculum, can hardly be exceeded in importance by any other skill-area, except possibly listening comprehension.

Thus, there is a special significance that attaches to the role of reading in the ELT curriculum at an academic institution of any level. The improvement of students' reading ability is an unending responsibility, so statements of curricular objectives, selection of course teaching materials, and content specifications of major course examinations generally reflect that fact. Nevertheless, it is not uncommon to find programs in which the explicit teaching and testing of reading comprehension are being overlooked in course syllabi. The usual explanation to be heard, even in EFL programs, is:

"We're not teaching reading in this course; therefore, we don't test it." This claim may even be made when 1) students are assigned challenging textbooks for such courses, comprehension of which may be essential in order to pass, and 2) when there are challenging reading passages in the major examinations as well.

One consequence of this state of affairs is that students may be held responsible for understanding their course materials, with insufficient help being given to them when they have serious comprehension problems. Such difficulties may stem from any combination of problems involving content, language and culture. Without regular check-ups throughout a course on reading comprehension, students may not be served well by their program, especially those many individuals in every course who are below-average in ability and require more personal attention than their classmates. Without major test tasks assessing reading ability on a regular basis, a staff will lack basic data about how well the students are progressing in this vital skill-area.

B. Advantages of Testing Reading
vis-a-vis the Other Skills

There are special advantages that attach to having, in test batteries, tasks validly assessing reading comprehension. Of the four macro-skill areas of language, reading can be most readily and efficiently tested in large ELT programs. For example, the testing of listening may require the use of special equipment and may give rise to worries about whether all testees can perceive the stimulus material equally well. Listening scores are also affected fundamentally by the troublesome variable of phonology.

As for the productive skills, these are tested most validly in direct-performance proficiency tests which are notoriously difficult to score reliably, even with trained raters. Such tests, while easy to set, are often very complicated to administer, as with tasks involving speaking (e.g., interviews) or oral exchanges (e.g., role plays); or complicated to mark, as with extended written discourse (e.g., essays, research reports, lab reports, etc.).

On the other hand, while good reading tests are among the most difficult of all tests to prepare, they are among the easiest, when objectively scored, to mark efficiently and reliably after the test adminis-

tration. Thus, when well-designed and -written, tests of reading---when compared with those of any of the other three macro-skill areas---tend often to produce the most valid scores of integrative language use and of general language proficiency, other things being equal.

III. The Validity of Reading Tests

A. Content and Construct Validity

A reading test is valid when it satisfies two important conditions: 1) that the scores it produces are primarily measurements of students' reading ability, and are not being critically influenced by other, mostly irrelevant, factors; and 2) that, since the main purpose of FL or SL tests is to produce differences in scores reflective of real differences in students' language abilities (Krzanowski and Woods 1984:3), the reading scores do distinguish with a reasonable degree of accuracy among students of markedly different levels of reading proficiency. The first condition concerns test validity and the second, reliability; and without the latter condition, the former cannot be satisfied. In the first constraint is emphasized the requirement that the test scores be defensible primarily as reading scores, not as something else "contaminating"

the test results (Malcolm, personal communication). In the second, it is the need for consistency among questions and tests to discriminate effectively and consistently between learners of clearly different ability or achievement levels.

In the above discussion, it is the word scores which must be highlighted. A class of students may be required to do a good deal of reading in a test part so labeled; but unless that part has been properly designed and written, the results may very well not be considered acceptable reading scores. The task(s) may appear to the teachers to have face validity, to constitute an appropriate sample of what students have been reading during their course (and may thus have content validity); but the test will lack construct validity---i.e., it will not be acceptable in its results to internationally recognized testing experts as an appropriate measure of reading ability.

It is in this sense that Davies (1986) says that "many ESP tests are samples but not measures"; that is, they lack score validity. As Farhady and Hatch point out, validity resides primarily not in the test itself---i.e., on its face--- but rather in its outcomes ---its scores---and in how those scores are then used (Hatch and Farhady, 1982:251). Thus, teachers may label a test part "Reading Comprehension," when in fact the

results do not technically qualify to be considered as true reading scores (Stevenson 1981: 270).

B. Test Design

The vital step for classroom teachers hoping to produce any kind of valid test is taken at the design stage. As Davies points out (1984a: 51), the key consideration in assuring test validity lies in how a test is designed. In tests of reading, the crucial elements are: 1) the content of the reading task(s), 2) the length of the stimulus material, 3) the question format(s), 4) the time allotted for completion of the task(s), and 5) the amount of information obtained about the students' reading ability. The effects of these several influences upon students' reading scores are closely intertwined and cannot be easily separated in evaluating test outcomes.

B.J. Carroll points out (B.J. Carroll 1980: 13) that a prime characteristic of a good test is economy: it produces a lot of information about a subject or group of subjects within a relatively limited period of time. Thus, in a good reading test, the stress for test developers, perhaps paradoxically, must be not on making students read too much, but rather on making them respond, as much as possible. Therefore, there is a premium to be placed on a reading-test design maximizing

the ratio of questions asked to the number of words read (Malcolm 1980).

A second characteristic of a suitable reading test is good sampling reliability (Malcolm, *ibid.*); that is, including a variety of topics, together with the differing lexis and discourse patterns associated with each topic and/or register. Thus, no students are given an inadvertent linguistic advantage by happening to be familiar with the particular subjects found in the passages presented.

To achieve the above two design features, a suitable reading test presents a variety of short reading passages with enough substance in them to produce at least one good question for every 20-25 words read. In a major achievement test, like a final examination, 30 good questions for 5-6 passages on quite different topics can be timed---for the majority of average-ability readers at that course level who have diligently prepared for their test---to be completed in 30-40 minutes. By emphasizing brevity in the passages, the testers do not force students to waste valuable exam time hunting around for references in passages or having to rely on memory. Difficulty is defined not by length of passages or number of questions but by the intrinsic challenge of the test content and language itself. And

of course there should be no comprehension problems in the language of the questions.

C. Question Format

1. productive vs. receptive integrative skills

One of the most crucial influences on students' test scores---and thus on test validity---is the impact of the format of questions. In the case of the productive macro-skills of speaking and writing, it is generally agreed that the most valid test formats require students to "produce": to speak and to write (Clark 1978:24). The main concern is to have multiple scorers of the language produced and to assure that they are able to apply the same scoring criteria in the same way---i.e., consistently, both 1) at the same time and 2) at different times for the same test tasks.

But whereas it is possible for scorers to record and examine (i.e., to perceive) the language produced by students, there is no direct way of perceiving what psycholinguistic phenomena take place inside the students' heads when they are hearing or reading. The concern of the examiner then becomes how to obtain test scores for these two vital skill-areas which are not influenced by such irrelevant factors as examiners' reactions to students' handwriting samples, spelling

errors, control of grammatical structure, control of writing mechanics, or ability to write (as an integrative skill).

2. "communicativeness" and practicality in tests

Consequently, while content-valid test tasks assessing the productive skills tend to be marked according to subjective criteria, the receptive skills are overwhelmingly tested by measurement authorities with objectively scored formats. Claims of many ELT practitioners "in the field" that there is abroad today a new communicative paradigm casting doubt on the suitability of objective tests are not borne out by inspection of the most widely-used standardized tests (Davies 1984b: 82). Most such validated tests employ objective formats to test listening and reading comprehension, usually the 4-option multiple-choice format.

The strengths and limitations of the 4MC format are discussed at length in the professional literature and need not be elaborated on here. However, given the indirect nature of assessing the receptive skills, it is worth asking what it means to call a listening or reading test "communicative." While the virtues of the communicative movement over the past decade have been significant in the improvement of teaching methodology and of classroom materials, and even of many aspects of testing, there are many ELT specialists who question the applicability of the adjective "communicative" to such words

"in the ELT lexicon (as) syllabus, materials, course design, testing. I believe it is misleading to do this, for these aspects of the organization of ELT are inherently non-communicative, except in the simple sense of being written for someone to understand and act upon. But the way these things get talked about, it is almost as if by having a 'communicative' syllabus, using 'communicative' materials, that a 'communicative' methodology is assured and 'communicative competence' is achieved, and that we can measure the achievement by a 'communicative' testing instrument" (Mountford 4: 1985).

The classroom is an artificial environment in many obvious respects. And within the classroom, the testing session presents even more rigid constraints on the natural use of the language in any of its modes, whether considered in combination or separately. In fact, the test itself---in ELT or in nearly any other academic discipline---isn't intended to duplicate reality because it can't, but rather to sample it scientifically, particularly in this age of mass education and large-scale testing in sizable educational institutions. The tester is often constrained in at least two vital respects, because of the magnitude and difficulty of the FL testing task: 1) in the extent of proficiency that can be assessed, because of the number of testees; and 2) in the amount of language that can be sampled because of the limits of time.

These two practical constraints, time and the number of examinees, have no doubt encouraged the increasing use of objective testing in large ELT programs worldwide in recent decades. On the positive side, a great deal of information about students can be obtained

efficiently and reliably within a very short time. And when the tests are well-designed and -written, teachers usually find not only that scores tend to be fair but that students are generally ranked accurately with respect to their relative levels of overall proficiency and/or achievement. As a consequence, classroom teachers and program administrators are usually confident in their use of the test scores, in the subsequent major program decisions they must make regarding admissions, placement, grading, promotion and graduation of students.

3. "the best" method of testing reading in ELT programs

Assuming that 4MC reading tasks are well-designed and the questions well-written, most testing experts seem to advocate this format for testing reading comprehension in ELT programs of academic institutions. Davies calls it "the best" method for assessing reading ability in both EAP and ESP programs (Davies, personal communication).

The reasons are not difficult to understand. First, the language presented is truly integrative, not discrete-point like that in sentence-level 4MC items testing vocabulary, structure or phonology. Second, the language presented is authentic written English discourse, which is especially "communicative" when that language is the published product of professional writers, unlike the "homemade" products that are

used in many ELT programs. Third, when passages are chosen for appropriateness of content, difficulty level and length, students tend to be very well ranked in their scores as to relative proficiency levels. Fourth, although the act of putting a mark in a box may appear to be an artificial one, the fact that the response mode is essentially a non-linguistic one prevents the intrusion of linguistic factors other than reading ability from influencing students' scores. Fifth, when at least 20 reading questions---and preferably more---are asked, an adequate data base is provided both for meaningful discrimination in part scores among students and for calculating statistics for item-bank purposes (Malcolm, personal communication).

There is nothing to prevent teachers, in a realistically timed test, to employ alternative formats for testing reading, either separately or in combination with other linguistic and/or behavioral variables (J.D. Carroll 1961:319-20), to see which demonstrate the greatest end-of-course concurrent validity (Bruton 1984: 117), a kind of criterion-related validity. They can then see themselves as experimenters (Allwright, in Shaw, et.al., 1983: 132), testing their assumptions to see which of the alternative formats rank students best in reading, in comparison with key course criteria like total semester marks, course grades, and their own intuitions and knowledge about their students.

Conclusion

In conclusion, accountability of EFL and ESL teachers is twofold: first, to teach their students as well as they can; and, second, to assess students' proficiency and progress as accurately as they can. The equal importance of these two sets of classroom activities is often obscured by the enormous difference in the amount of time devoted to each during a course. Students' regard for their language instructors is first and foremost influenced by how well they believe they have been taught and how much they think they have learned. But the satisfaction derived from much excellent instruction can quickly be undermined if students, especially as a group, have legitimate reasons to doubt the effectiveness, comprehensiveness, stringency or accuracy of post-instruction assessment efforts.

Teachers serve themselves well by knowing how to continually improve both their teaching and their testing methods. The decision to use or not to use particular testing formats is best informed by knowing what the foremost authorities in the field recommend in a particular case, and why. The result may not be the instant solution to all testing problems, but such an approach to dealing with those problems should gradually help to put the entire

decision-making process regarding testing policy and practice in ELT programs into clearer perspective and onto a much surer footing.

References

- Alderson, J.C. and A. Hughes. 1981. Issues in Language Testing: ELT Documents No. 111. London: The British Council.
- Bruton, C. 1984. Review of A Language Testing Handbook, by A. Harrison (1983). *Language Testing* 1:1.
- Carroll, B.J. 1980. Testing Communicative Performance. Oxford: Pergamon Press.
- Carroll, J.D. 1972. Fundamental Considerations in Testing for English Language Proficiency of Foreign Students. In Teaching English as a Second Language (H.B. Allen & R.N. Campbell, eds.). New York: McGraw-Hill International Book Company.
- Clark, J.L.D. 1978. Psychometric Considerations in Language Testing: Advances in Language Testing Series 2 (B. Spolsky, ed.). Arlington, Va.: Center for Applied Linguistics.
- Cohen, A. 1980. Testing Language Ability in the Classroom. Rowley, Mass.: Newbury House Publishers, Inc.
- Davies, A. 1984a. Validating Three Tests of English Language Proficiency. *Language Testing* 1:1.
- , 1984b. ESL Expectations in Examining: The Problem of English as a Foreign Language and English as a Mother Tongue. *Language Testing* 1:1.
- , 1986. Raising Plateaus of Proficiency. Paper presented at IATEFL conference in Brighton, England, April 1986.

- Hatch, E. and H. Farhady. 1982. Research Design and Statistics for Applied Linguistics. Rowley, Mass.: Newbury House Publishers, Inc.
- Henning, G. 1982. Twenty Common Testing Mistakes for EFL Teachers to Avoid. *English Teaching Forum* 20:3.
- Krzanowski, W.J. and A.J. Woods. 1984. Statistical Aspects of Reliability in Language Testing. *Language Testing* 1:1.
- Malcolm, D.J. 1980. *Test Development Manual*. Kuwait University Language Center.
- Mountford, A. 1985. Practising Communicative Language Teaching in Thailand. Paper presented at Thai/ TESOL convention in January 1985, Bangkok.
- Nakhoul, E. 1985. The Best of Both Worlds: An Attempt to Reconcile Some Dichotomies in Current Testing Theory in the Design of a Placement Test System. In Proceedings of the First National Symposium on Language Teaching (Vol. II), May 1985, Kuwait.
- Orem, R. 1981. Entering the '80s---Some Professional Perspectives. *TESOL Newsletter* 15:2.
- Palmer, J.D. 1983. What Makes a Language Teacher a Professional? *ESP Journal* 2:1.
- Shaw, P., et. al. 1983. Talking Shop: The L.A. Tapes. *ELT Journal* 37:2.

Stevenson, D.K. 1981. "All of the Above": On
Problems in the Testing of FL Reading. System
9:3.

Woodford, P.E. 1980. Foreign Language Testing.
The Modern Language Journal 64:1.